Calculating indices for urban sprawl

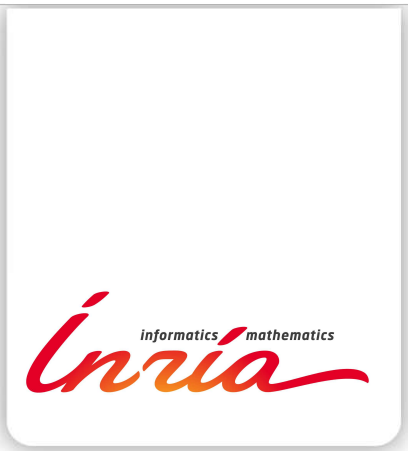
Jakub Krzywda, Peter Sturm

Calculating indices for urban sprawl

Jakub Krzywda, Peter Sturm

RESEARCH REPORT
N° 8398
November 2013
Project-Team STEEP

ISSN 0249-6399 ISRN INRIA/RR--8398--FR+ENG



Calculating indices for urban sprawl

Jakub Krzywda*, Peter Sturm

Project-Team STEEP

Research Report n° 8398 — November 2013 — 44 pages

Abstract: Urban sprawl is a complex concept [8], that is generally associated with auto-oriented, low-density development. It is the subject of a wide range of research efforts, aiming at understanding and characterizing the underlying driving factors. This report follows an effort by Burchfield et al. who proposed in [1] a simple measure for urban sprawl, a so-called sprawl index. It proposes variants of this index and describes their implementation using the R statistical computation environment [2], the Geospatial Data Abstraction Library [7] and the Quantum GIS (Geographic Information System) [5].

Key-words: Urban sprawl, sprawl index

* This work was done within the INRIA international internships programme.

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Calcul d'indexes d'étalement urbain

Résumé : L'étalement urbain est un concept complexe [8] qui est généralement associé à un développement de faible densité et basé sur l'automobile. Il fait l'objet de beaucoup de recherches ayant pour but la compréhension et la caractérisation des facteurs sous-jacents. Ce rapport suit un travail par Burchfield et al. qui ont proposé, dans [1], une mesure simple pour l'étalement urbain, appelé indexe d'étalement. Nous proposons des variantes pour cet indexe et décrivons leur implémentation utilisant l'environnement de calcul statistique R [2], la *Geospatial Data Abstraction Library* [7] et le SIG (Système d'Information Géographique) Quantum [5].

Mots-clés : Étalement urbain, indexe d'étalement

Contents

1 Introduction

Urban sprawl is a complex concept [8], that is generally associated with auto-oriented, low-density development. It is the subject of a wide range of research efforts, aiming at understanding and characterizing the underlying driving factors. This report follows an effort by Burchfield et al. who proposed in [1] a simple measure for urban sprawl, a so-called sprawl index. It proposes variants of this index and describes their implementation using the R statistical computation environment [2], the Geospatial Data Abstraction Library [7] and the Quantum GIS (Geographic Information System) [5].

In [1] as well as in this report, sprawl indices are computed over some of the largest Metropolitan Statistical Areas (MSA) of the United States of America. The computation is mainly based on land cover data, e.g. on the main usage of raster cells of land (residential, industrial, agriculture, etc.) and related density information.

The report is organized as fallows. Section 2 describes input data used. Section 3 explains data preprocessing steps. Section 4 presents the original sprawl index definition and computation method of [1], while in Section 5 newly proposed methods are described. Section 6 discusses various methods of comparing histograms and in Section 7 the selected ones are applied to cluster similar urban areas. Section 8 provides a summary and indicates possible directions for future work.

2 Input data

To calculate a sprawl index of each Metropolitan Statistical Area (MSA), two input files are used: a raster file with land cover information and a shapefile with the boundaries of MSAs. We have used essentially the same data as [1]; these data as well as where to obtain them, are described extensively in [1] and on `http://diegopuga.org/data/sprawl/`. In the following, we only describe the main characteristics of the data used, please refer to the above references for more details.

2.1 Land cover raster file

The land cover raster file contains information on the land cover of the whole contiguous United States. The surface is divided into about 8.7 billion cells of size $30 \times 30m^2$. Each cell has information about its land cover.

During our research, three datasets (raster files) from different years were used. The oldest dataset describes land cover from 1992, the second one from 2001 and the newest one from 2006. Figure 1 shows the land cover information from 2006.

The land cover classification system was changed between 1992 and the later years under consideration. The following list describes the different means of classification of developed land in the National Land Cover Datasets (NLCDs) used.

NLCD1992 distinguishes two types of development: residential and other (including commercial, industrial and transportation). Residential development is divided into two levels of intensity: low (constructed materials account for 30% to 79% of cover) and high (80% to 100% of constructed materials).

NLCD2001, NLCD2006 do not split development by purpose. However, they distinguish four levels of intensity: open space (0% to 19% of impervious surfaces), low (20% to 49%), medium (50% to 79%) and high intensity (80% do 100%).

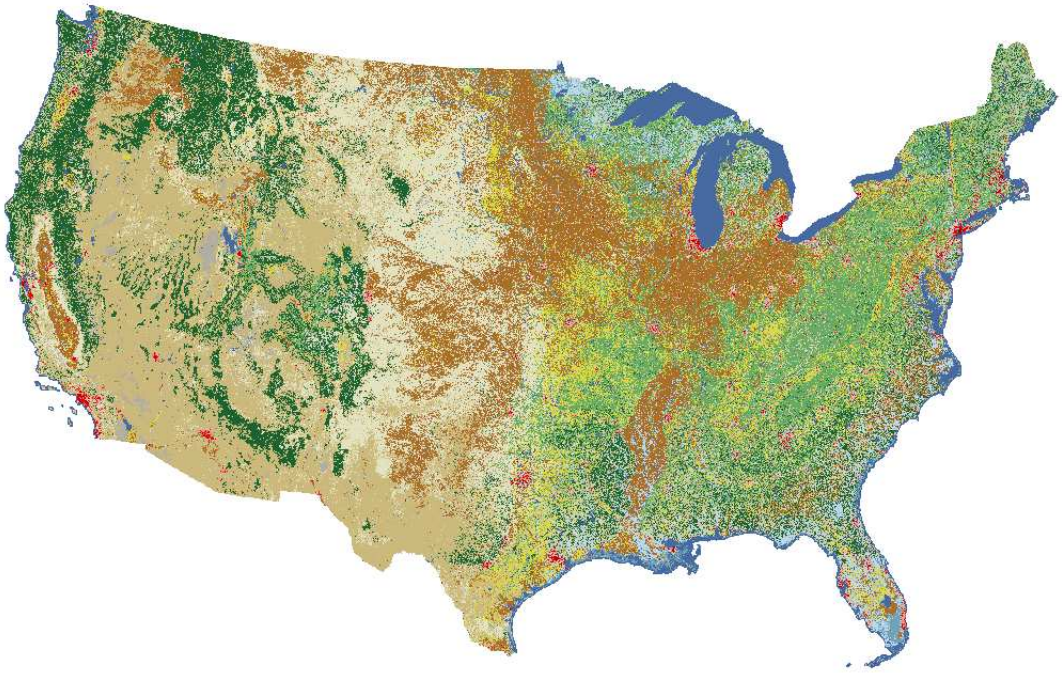


Figure 1: Raster file with land cover for the year 2006.

Table 1: Differences in classification systems used in raster files.

	NLCD1992	NLCD2001, NLCD2006
Number of categories	9	8
Number of subcategories	21	20
Number of development intensity levels	2	4
Separate subcategories for residential land	Yes	No

Table 1 summarizes differences between classification systems, while Appendices A and B present the entire hierarchy of land cover categories.

2.2 Shapefile

A shapefile contains geospatial data, usually in a vector format. In general, it is able to store coordinates of points, lines and polygons together with associated data. In case of our application, a shapefile contains the set of polygons representing boundaries of MSAs with attributes specifying name and code. A more detailed specification of the used ESRI (Environmental Systems Research Institute) shapefile can be found in [3]. Figure 2 shows the boundaries of all MSAs located in the contiguous United States.

To properly extract land cover of MSAs from the raster file, both input files must have the same projection, i.e. one must be able to align land cover raster cells with the MSA boundaries from the shapefile. During preprocessing, the projection of a shapefile can be changed using a GIS software.

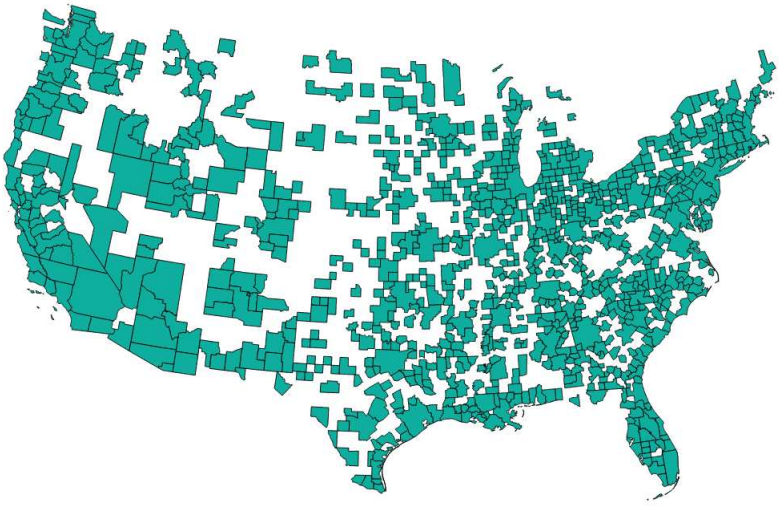


Figure 2: Shapefile with boundaries of MSAs

3 Data preprocessing

Data must be properly prepared before computing sprawl indeces. In our case, preprocessing consists of four steps: selection and separation of MSA boundaries; buffering boundaries of MSAs; extraction of raster data; and grouping of land cover categories.

3.1 Separation of MSA boundaries

The shapefile contains boundaries of all MSAs but in the next steps the boundary of just one MSA at a time will be needed. Because of that it is necessary to select the boundary of the MSA which will be processed, separate it from the others and export it to a new intermediate file `msa_boundary`. The selection of an MSA can be based on its code (numeric value) or its name (textual value).

An MSA boundary may consist of more than one polygon. It occurs mostly on the coasts where some MSAs includes both continental land and islands. Figure 3 shows the Santa Barbara MSA in Southern California which consists of one area on the mainland and three islands.

Listing 1 shows the associated invocation of the `ogr2ogr` program. It selects polygons from file `shapefile` using the code of the MSA specified in `msa_code`. The next program saves them in the file `msa_boundary`.

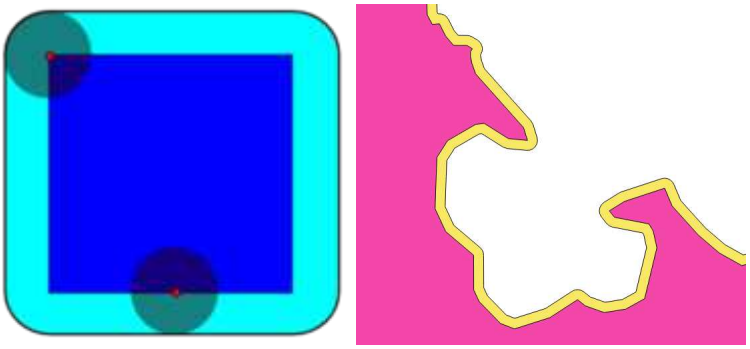
Listing 1: Separation of MSA boundaries

```
ogr2ogr [msa_boundary] [shapefile] -where "CBSA = '[msa_code]'"
```

3.2 Buffering MSA boundaries

Sprawl indices are based on measuring characteristics of surroundings of individual raster cells, where the surroundings is defined for instance by a distance (a radius). Problems occur near boundaries of MSAs, when that distance is larger than the distance of raster cells to the MSA

Figure 3: Boundaries of the Santa Barbara MSA.



(a) Dilation of a square by a circle. (b) Buffering of a part of the Boston MSA.

Figure 4: Dilation and buffering.

boundary. During calculations, cells with surroundings going beyond boundaries of the processed raster take a value of "NA" (not available). The sprawl index will not include the whole area when the extracted raster has the same size as the boundaries of the considered MSA. Polygons separated in the previous step must be properly enlarged to ensure that the sprawl index describes the entire area, including strips of land on the boundary.

The process of enlarging polygons in GIS software is called buffering. It corresponds to a dilation in mathematical morphology. A buffer is the set of all cells around an area within the specified maximum distance.

Figure 4 shows the dilation of the dark-blue square by a circle, resulting in the light-blue square with rounded corners (a) and the buffering of the pink polygon representing a fragment of the Boston MSA, resulting in the yellow polygon (b).

Listing 2 shows a Python script responsible for buffering boundaries of MSAs. Function `QgsVectorLayer()` loads polygons from the file `msa_boundary`. Parameters `layer_name` and `provider` specify the name used to represent the layer and the name of the data provider respectively. The returned MSA boundary is assigned to variable `layer`. Function `QgsGeometryAnalyzer().buffer()` enlarges `layer` by adding a buffer having a width of `buffer_size` meters and saves it in file `msa_buffered`.

Listing 2: Buffering boundaries of MSA

```
layer = QgsVectorLayer("[msa_boundary]",
                       "[layer_name]",
                       "[provider]")

QgsGeometryAnalyzer().buffer(layer,
                             "[msa_buffered]",
                             [buffer_size],
                             False,
                             False,
                             -1)
```

3.3 Extraction of MSA land cover

In the next step of preprocessing, the land cover information for an MSA is extracted from the input raster using the previously obtained polygons. Listing 3 shows the associated invocation of the `gdalwarp` program. It extracts from `raster_file` the land cover information of the area described by polygons in `msa_buffered` and saves it to file `msa_raster`.

Listing 3: Extraction of MSA land cover

```
gdalwarp -dstnodata 0 -cutline "[msa_buffered]"
         -crop_to_cutline -of HFA [raster_file] [msa_raster]
```

3.4 Grouping of land cover categories

The input raster file contains 20 categories of land cover (21 in the NLCD1992 version) including different levels of development intensity. This is more than needed for the calculation of the sprawl indices considered in this report. In order to simplify computations, a reclassification is made. Categories of land cover are grouped into three general classes: developed, undeveloped and water. Because the original sprawl index of [1] is calculated only across residential land, a fourth class is also separated: the set of residential cells is a subset of the developed cells.

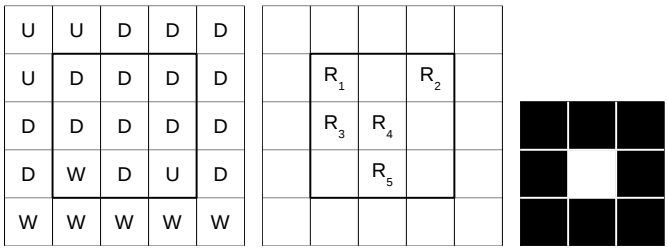
Listing 4 shows part of a script in the R language. Firstly, data are loaded from file `msa_raster` to variable `landCover`. Next, based on a specified matrix, a reclassification is made: value 11 (water) becomes 1, values in the range 20–29 (developed) become 2, the rest (undeveloped) becomes 0. The result is assigned to the new variable `devUndevAndWater`. An auxiliary raster `residential` is created with value 1 for residential land and "NA" for others.

Listing 4: Grouping of land cover categories

```
landCover <- raster([msa_raster])
devUndevAndWater <- reclassify(landCover, c(0,10,0, 10,11,1, 11,19,0, 19,29,2, 29,
    Inf,0))
residential <- reclassify(raster, c(0,20,0, 20,22,1, 22,Inf,0))
residential[residential == 0] <- NA
```

4 Sprawl index calculation

This section describes the main computations of the sprawl index as it is defined in [1]. Firstly, a simple example is presented. Next, a formal definition of the sprawl index is given. Last, the most important fragments of source code are explained.



(a) `devUndevAndWater` (b) `residential` (c) `mask`

Figure 5: Input matrices (a), (b) and surrounding mask (c); `D` – developed, `U` – undeveloped, `W` – water, `R` – residential.

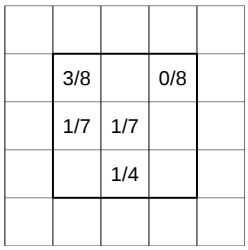


Figure 6: Matrix `surroundingRatio` with the proportions between undeveloped cells and the total number of dry land cells in the surroundings of each residential cell.

4.1 A simple example

This example shows the different steps of the sprawl index calculation. It is assumed that during the preprocessing phase the input data are converted into two matrices containing all necessary data, as described in Section 3.4.

Figure 5 presents these precomputed data. The matrix `devUndevAndWater` with information about land cover categories is shown in (a). The sprawl index will be calculated in the central area A of size 3×3 cells limited by the bold line. The location of residential areas is presented in (b). The surrounding mask shown in (c) includes all neighboring cells with a Chebyshev distance lower or equal to 1, so it is a square of 3×3 cells, the center being excluded. The buffer of the area must be included in computations to obtain the sprawl index for the whole area, as mentioned in Section 3.2. The width of the buffer is equal to the radius of the surrounding, so the entire buffered area has a size of 5×5 cells.

The **surrounding ratio** $\sigma(c)$ shows the proportion between undeveloped cells and all dry land cells in the surroundings of cell c. In the example, there are three undeveloped cells out of eight dry land cells in the surroundings of R_1, so $\sigma(R_1) = 3/8$. For R_2, none of surrounding cells is undeveloped, therefore $\sigma(R_2) = 0/8 = 0$. For both R_3 and R_4, there is one cell with water and one undeveloped cell in their surroundings, so $\sigma(R_3) = \sigma(R_4) = 1/7$. The surroundings of R_5 contain only four dry land cells and only one of it is undeveloped, therefore $\sigma(R_5) = 1/4$. The values of the surrounding ratios of each residential cell are shown in Figure 6.

Finally, the sprawl index of area A, is the average of the surrounding ratios over all residential cells inside A:

$$SI(A) = \frac{3/8 + 0/8 + 1/7 + 1/7 + 1/4}{5} \approx 0.18$$

4.2 Mathematical description

The original sprawl index shows the ratio of undeveloped cells in the surroundings of an average residential cell.

Function $\delta(c)$ shows if cell c is classified as developed:

$$\delta(c) = \begin{cases} 1 & \text{if } c \text{ is developed,} \\ 0 & \text{if } c \text{ is undeveloped.} \end{cases}$$

Function $\rho(c)$ shows if cell c is classified as residential:

$$\rho(c) = \begin{cases} 1 & \text{if } c \text{ is residential,} \\ 0 & \text{otherwise.} \end{cases}$$

Function $\omega(c)$ shows if cell c is classified as water or dry land:

$$\omega(c) = \begin{cases} 1 & \text{if } c \text{ is open water,} \\ 0 & \text{otherwise.} \end{cases}$$

R_A is the set of all residential cells in area A:

$$R_A = \{x : x \in A \wedge \rho(x) = 1\}$$

S_c is the surroundings of cell c. In [1], it consists of all dry land cells completely contained within a square of one kilometer side length, whose center is in the middle of cell c:

$$S_c = \{x : d_{ch}(x, c) \leq 500\text{m} \wedge \omega(x) = 0\}$$

where d_{ch} is the Chebyshev distance or L_∞, $d_{ch}(p, q) = max_i(|p_i - q_i|)$.

The surrounding ratio $\sigma(c)$ is the number of undeveloped cells divided by the number of all cells in the surroundings of cell c:

$$\sigma(c) = \frac{\sum_{i \in S_c}(1 - \delta(i))}{|S_c|}$$

Finally, the sprawl index of area A is the average value of surrounding ratios for all residential cells within A:

$$SI(A) = \frac{\sum_{c \in R_A} \sigma(c)}{|R_A|}$$

4.3 Source code explanation

In this section the most important parts of source code for the computation of the sprawl index are explained.

Listing 5 shows a script computing the sprawl index. The second line of the listing is the most important. Function `focal()` calculates for each cell in the area `devUndevAndWater` its surrounding ratio. The ratio is determined by applying function `focalFun` on the surroundings specified by `mask`. Function `focalFun` returns the ratio between the number of developed cells and the number of dry land cells in the surroundings. The result of the `focal()` function is stored in variable `surroundingRatio`. In the next step values from the `surroundingRatio`

matrix are multiplied by values from the `residential` matrix and saved in variable `sprawl`. This multiplication causes that in the result matrix `sprawl` only values of surrounding ratios from residential cells are kept. To calculate the final value `sprawlIndex` all values stored in the `surroundingRatio` matrix are summed up and divided by the number of residential cells. The sum is computed using the `cellStats()` function while the number of residential cells is computed using function `freq()`.

Listing 5: Sprawl index calculation

```
focalFun <- function(x) { return(sum(x == 0) / sum(x != 1)) }
surroundingRatio <- focal(devUndevAndWater, w=mask, fun=focalFun)
sprawl <- surroundingRatio * residential
residentialCells <- freq(residential, value=1)
sprawlIndex <- cellStats(sprawl, stat="sum") / residentialCells
```

A sample definition of the surrounding mask is shown in Listing 6. For simplicity, the presented surrounding mask has a small size of 5×5 cells. The size used in actual computations is 30×30 cells.

Listing 6: Sample definition of a surrounding mask

```
mask = matrix(c(1,1,1,1,1,
                1,1,1,1,1,
                1,1,0,1,1,
                1,1,1,1,1,
                1,1,1,1,1), nrow=5)
```

5 Extensions for the sprawl index

This section describes newly proposed methods for measuring area scatteredness, extending the original sprawl index of [1]. Three scalar measures – generalized sprawl index, inverted sprawl index, and intensity sprawl index – are presented and compared. Next, the interval sprawl index is introduced, that takes into account an uncertainty of development intensity measurement. At the end of this section, a new approach to represent area scatteredness by using histograms is described.

5.1 Generalized sprawl index

Because current versions of NLCD (2001 and later) do not contain information about the purpose of development it was decided to calculate the sprawl index across all developed cells.

D_A is the set containing all developed cells in area A:

$$D_A = \{x : x \in A \wedge \delta(x) = 1\}$$

The generalized sprawl index shows the percentage of undeveloped cells in the surroundings of an average developed cell:

$$GSI(A) = \frac{\sum_{c \in D_A} \sigma(c)}{|D_A|}$$

5.2 Inverted sprawl index

The inverted sprawl index shows the percentage of developed cells in the surroundings of an average developed cell. It is an intermediate measure, geared towards a more complex one proposed below.

$$\sigma_{inv}(c) = \frac{\sum_{i \in S_c} \delta(i)}{|S_c|}$$

5.3 Intensity sprawl index

The raster contains information about the intensity of development which was not used by the previous measures. To take into account these data, a new version of sprawl index is proposed. The intensity sprawl index shows the average percentage of development in the surroundings of an average developed cell.

Function $\delta_{avg}(c)$ takes five values instead of two, as in the previous version. The intensity levels depend on the percentage of impervious surfaces covering a cell as mentioned in Section 2.1. The values of $\delta_{avg}(c)$ are equal to the centers of the percentage ranges reported in Section 2.1.

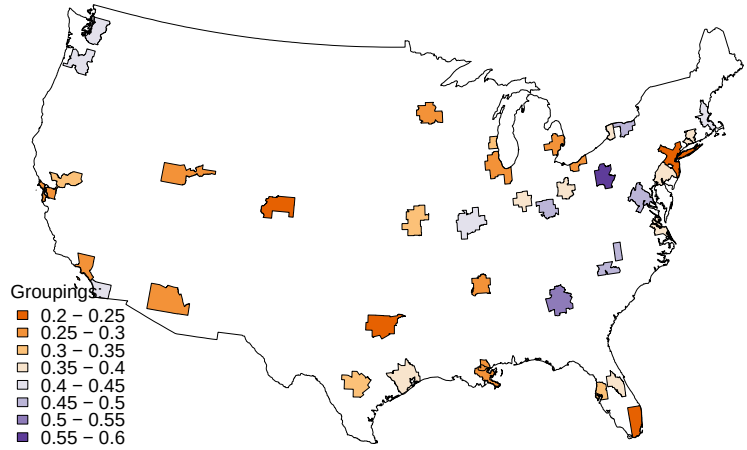
$$\delta_{avg}(c) = \begin{cases} 0 & \text{if } c \text{ is undeveloped,} \\ 0.1 & \text{if } c \text{ is open space,} \\ 0.35 & \text{if } c \text{ is low intensity,} \\ 0.65 & \text{if } c \text{ is medium intensity,} \\ 0.9 & \text{if } c \text{ is high intensity.} \end{cases}$$

For the computation of the intensity sprawl index the fourth step preprocessing – grouping of land cover categories – is omitted.

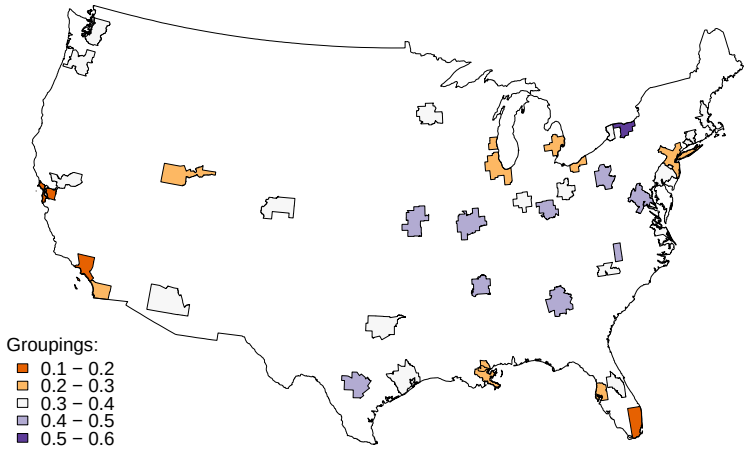
5.4 Comparison of scalar measures

Figure 7 shows maps of MSAs with associated values of the standard sprawl index for the dataset from 1992 and the generalized sprawl index for datasets from 2001 and 2006. Figure 8 presents the intensity sprawl index for all datasets.

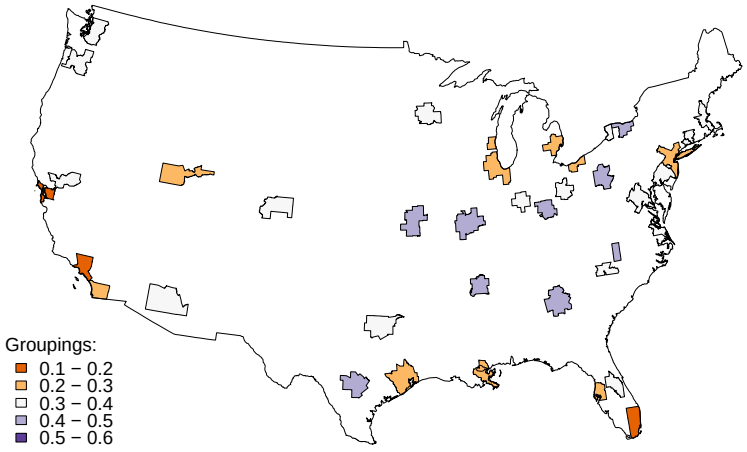
The ranges of values taken by the various indices in the six analyzed cases differ. First, the minimum values for cases from 1992 are higher for more recent years. This is caused by a different scale of development intensity levels (see Section 2.1). Moreover, the intensity sprawl index for datasets from 2001 and 2006 takes lower values than the generalized one. This is because the maximum value of the surrounding ratio is equal to the middle of the last interval, $max\left(\delta_{avg}(c)\right) = 0.9$.



(a) 1992 standard

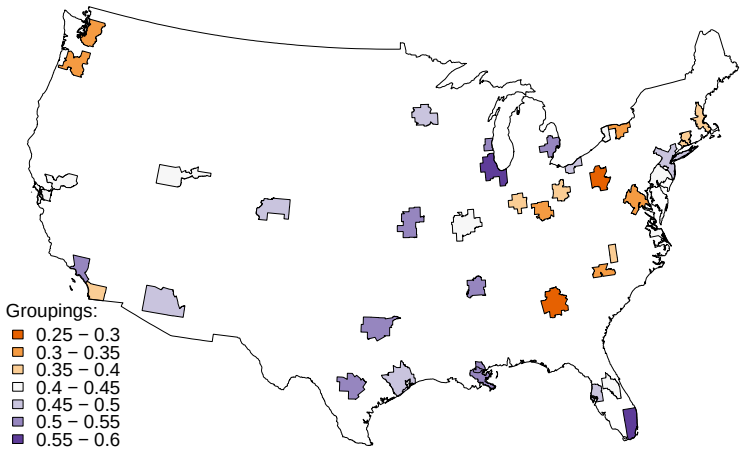


(b) 2001 generalized

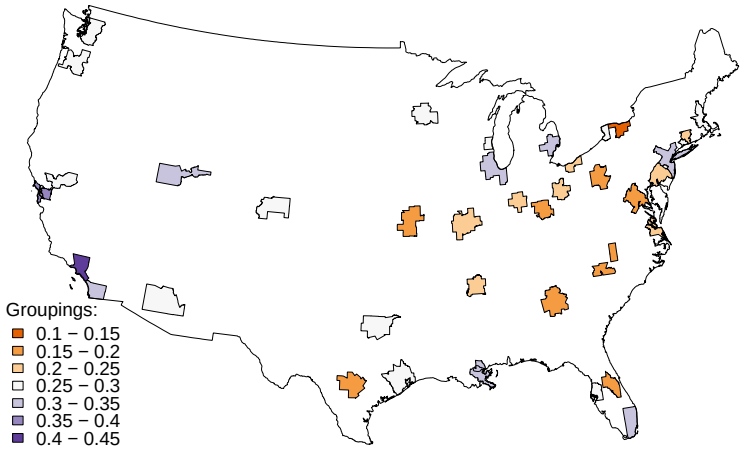


(c) 2006 generalized

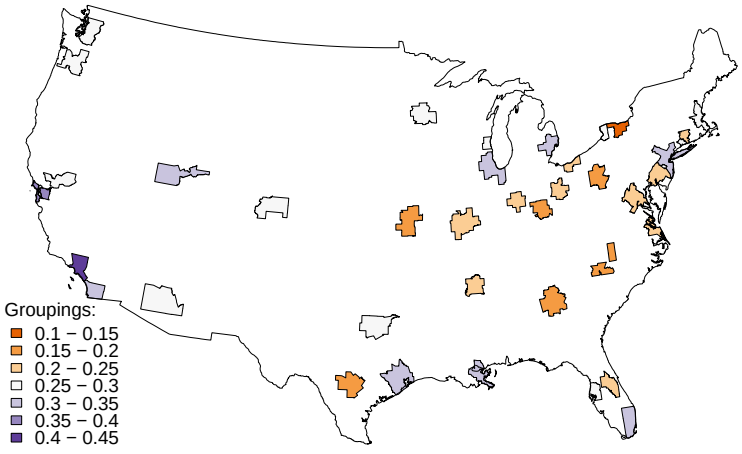
Figure 7: Visualization of standard (1992) and generalized (2001 and 2006) sprawl index.



(a) 1992 intensity



(b) 2001 intensity



(c) 2006 intensity

Figure 8: Visualization of intensity sprawl index measure.

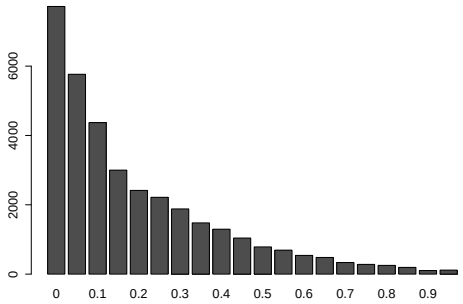


Figure 9: This histogram shows the distribution of surrounding ratios across the Miami region (MSA 33100). The height of each bin represents the number of cells with a value of surrounding ratio inside the corresponding 5% interval.

5.5 Interval sprawl index

As mentioned above, levels of development intensity are discretized in the raster file. Due to this, precise values of development intensity are not known. The only thing one can be sure of is that the value lies between the beginning and the end of the percentage range of the considered level.

To handle such an uncertainty, a new approach is proposed. The interval sprawl index shows the percentage range of development in the surroundings of an average developed cell. Its value is an interval and it is computed based on ranges for each cell.

The values of $\delta_{min}(c)$ and $\delta_{max}(c)$ are equal to the beginning (the lowest value) and the end (the highest value) of that percentage range respectively:

$$\delta_{min}(c) = \begin{cases} 0 & \text{if } c \text{ is open space} \\ 0.2 & \text{if } c \text{ is low intensity} \\ 0.5 & \text{if } c \text{ is medium intensity} \\ 0.8 & \text{if } c \text{ is high intensity} \end{cases}$$

$$\delta_{max}(c) = \begin{cases} 0.2 & \text{if } c \text{ is open space} \\ 0.5 & \text{if } c \text{ is low intensity} \\ 0.8 & \text{if } c \text{ is medium intensity} \\ 1 & \text{if } c \text{ is high intensity} \end{cases}$$

5.6 Histogram of surrounding ratios

Burchfield et al. proposed to use the arithmetic mean over all residential cells as a sprawl index value. The extensions described above use only a single mean value or an interval including that mean value. The arithmetic mean is a simple measure for a central tendency and using only it causes a loss of information, particularly when the distribution is not normal.

To describe a distribution of surrounding ratios across an area more precisely, a histogram is used. Figure 9 shows the distribution of surrounding ratios in the Miami region (MSA 33100).

6 Histogram dissimilarity measures

In order to compare two distributions of surrounding ratios, a dissimilarity measure is needed. A desirable measure should follow natural human perception. It should take into account sophisti-

cated relations between bins, not only local differences.

In this section six methods of measuring distances between histograms are presented. Firstly, examples based on bin-by-bin comparisons are described. Next, a cross-bin approach is shown. Finally, all measures are applied to a sample set of histograms and results are compared.

6.1 Bin-by-bin comparison

In a bin-by-bin approach only bins with equal indices are compared. The final dissimilarity measure is a combination of all pairwise comparisons.

Three measures are described: Minkowski-form distance, histogram intersection and χ^2 statistics.

Minkowski-form distance

$$d_{L_r}(H,K) = \left(\sum_i |h_i - k_i|^r\right)^{1/r}.$$

To compute the Minkowski-form measure absolute values of differences between corresponding bins of histograms are calculated. Differences are raised to the power of r and summed up. The final measure is the r^{th} root of that sum. Three forms are most commonly used: L_1, L_2 and L_∞. The last one can be presented as $d_{L_\infty} = max_i\left(|h_i - k_i|\right)$.

Histogram intersection

$$d_\cap(H,K) = 1 - \frac{\sum_i min\left(h_i, k_i\right)}{\sum_i k_i}.$$

Histogram intersection was proposed by Swain and Ballard in [6]. The value of this measure is inversely correlated with the sum of the heights of the smaller among two corresponding bins, over the whole range. In case of our application, the sum of each histogram is equal to one, so the denominator can be ignored. As shown by the authors of [6], when the areas of the two histograms are equal, what is always true in our case, the histogram intersection is equivalent to the normalized L_1 distance.

χ^2 statistics

$$d_{\chi^2}(H,K) = \sum_i \frac{\left(h_i - m_i\right)^2}{m_i},$$

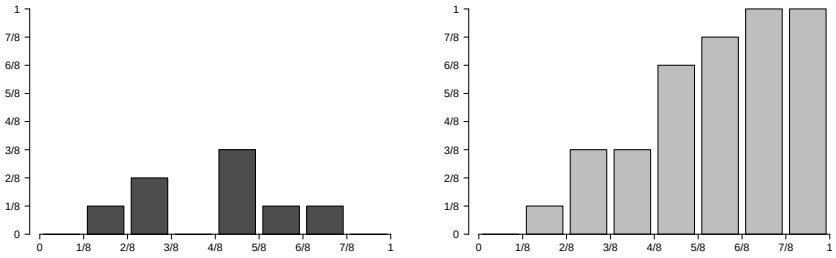
where $m_i = \frac{h_i + k_i}{2}$.

The χ^2 statistics measure refers to Pearson's χ^2 test that shows if the frequency distribution observed in a sample is consistent with a theoretical one. Here, two observed distributions of surrounding ratios are compared.

6.2 Cross-bin comparison

The cross-bin measures try to also take into account relationships between bins with different indices. To achieve this, several techniques can be used. Two measures, one using cumulative histograms and the other being based on a transportation problem approach, are described below.

Figure 10 shows a sample histogram and the corresponding cumulative histogram. The value of the i^{th} bin in the cumulative histogram is the sum of values of all bins of the original histogram



(a) Histogram (b) Cumulative histogram

Figure 10: Two different ways of presenting a distribution.

with indices less or equal to i. The cumulative histogram is always non-decreasing; the growth in each bin is equal to the value of the corresponding bin in the original histogram.

Formally, the cumulative histogram $\{\hat{h}_i\}$ of histogram $\{h_i\}$ is $\hat{h}_i = \sum_{j \leq i} h_j$.

Match distance

$$d_M(H, K) = \sum_i |\hat{h}_i - \hat{k}_i|.$$

The match distance between histograms is defined as the sum of differences in height of cumulative histogram bins. Therefore, the match distance can be also described as the L_1 distance between the cumulative histograms.

Kolmogorov-Smirnov distance

$$d_{KS}(H, K) = \max_i \left(|\hat{h}_i - \hat{k}_i| \right).$$

The Kolmogorov-Smirnov distance is a commonly used statistical measure to compare continuous probability distributions. Its value is equal to the maximum difference between cumulative histograms.

Earth mover's distance

Intuitively, one can see a distribution as a mass of earth spread across an N-dimensional space. The distance between two distributions is the least amount of work that must be done to move earth in the second distribution, in order to achieve a distribution equal to the first one. A unit of work corresponds to moving a unit of earth by a unit of ground distance.

In our case only a one dimensional space is considered. Let $P = \{(p_1, h_{p_1}), \ldots, (p_m, h_{p_m})\}$ be the first histogram with m bins, where p_i is the ground position and h_{p_i} is the height of the bin; $Q = \{(q_1, h_{q_1}), \ldots, (q_n, h_{q_n})\}$ the second histogram with n bins; and $D = [d_{ij}]$ the ground distance matrix where d_{ij} is the ground distance between bins p_i and q_j.

We want to find a flow $F = [f_{ij}]$, with f_{ij} the flow between p_i and q_j, that minimizes the overall cost

$$WORK(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij},$$

subject to the following constraints:

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq n \tag{1}$$

$$\sum_{j=1}^{n} f_{ij} \leq h_{p_i} \qquad 1 \leq i \leq m \tag{2}$$

$$\sum_{i=1}^{m} f_{ij} \leq h_{q_j} \qquad 1 \leq j \leq n \tag{3}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = min\left(\sum_{i=1}^{m} h_{p_i}, \sum_{j=1}^{n} h_{q_j}\right) \tag{4}$$

Once the transportation problem is solved and the optimal flow F is found, the earth mover's distance is defined as the resulting work normalized by the total flow:

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}.$$

It is worth emphasizing that a normalized value of the earth mover's distance is equal to a normalized value of the match distance.

6.3 Comparison of measures

To compare the above described measures, four sample histograms are used. Distances between pairs of histograms are calculated and the obtained values are confronted with intuitive human perception. The examples are based on those proposed in [4].

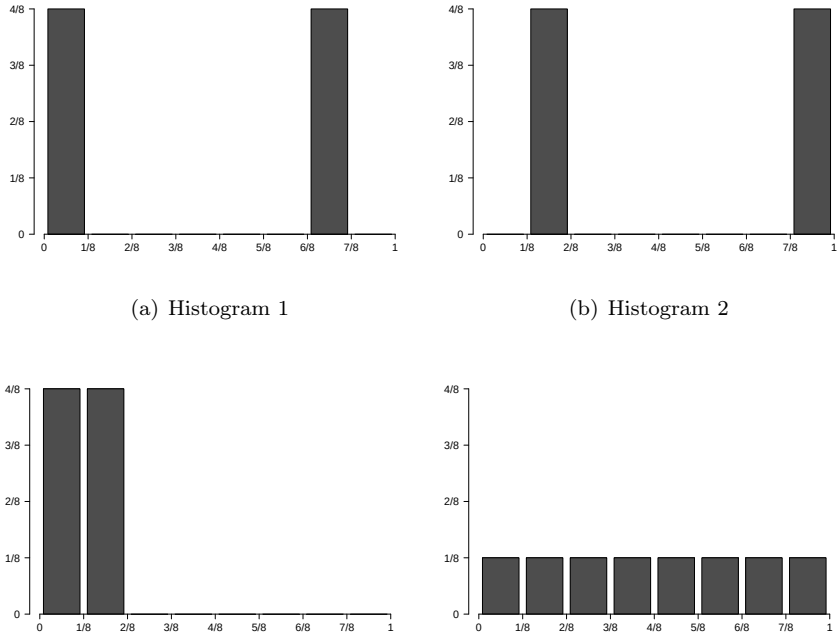
Figure 11 shows the sample histograms, consisting of eight bins each. The width of every bin is equal to 1/8 and the ground distance between bins is calculated between their centers. The only difference between histogram 1 (a) and histogram 2 (b) is a one-bin shift. Histogram 3 (c) consists, unlike the two previous ones, of two adjacent bins. The last one, histogram 4 (d), is totally different and presents a uniform distribution. Intuitively, histogram 1 is more similar to histogram 2 than to histogram 3 or histogram 4.

The desired measure should follow this intuition, so, more formally, it should satisfy the following conditions:

$$d(H1, H2) < d(H1, H3) \tag{5}$$

$$d(H1, H2) < d(H1, H4) \tag{6}$$

Table 2 shows distances between three pairs of histograms: $(H1, H2)$, $(H1, H3)$ and $(H2, H3)$. The values of seven distance measures are given for each pair. As mentioned before, values of histogram intersection are equal to those obtained from the normalized L_1 measure. Also, values of match distance and earth mover's distance are equal after normalization. In all these cases, normalization means division by the maximal possible value of the respective measure. The conditions given in Equation (5) and Equation (6) are met only by the match distance and the earth mover's distance.



(a) Histogram 1

(b) Histogram 2

(c) Histogram 3

(d) Histogram 4

Figure 11: Histograms used to compare dissimilarity measures: bimodal (a), bimodal shifted (b), unimodal (c), uniform (d). In accordance with human perception, histogram 1 is similar to histogram 2, but different from histograms 3 and 4. However, most of the analyzed histogram distance measures indicate the opposite.

Table 2: Comparison of histogram dissimilarity measures

	L_1	L_2	L_∞	HI	χ^2	MD	KS	EMD
d_{max} [1]	2.0	1.4142	1.000	1.00	1.0	7.00	1.000	0.8750
$d(H1, H2)$	2.0	1.0000	0.500	1.00	1.0	1.00	0.500	0.1250
$d(H1, H3)$	1.0	0.7071	0.500	0.50	0.5	2.50	0.500	0.3125
$d(H1, H4)$	1.5	0.6124	0.375	0.75	0.6	1.25	0.375	0.1562

Notes: L_1, L_2, L_∞ – Minkowski-form distances with $r = 1$, $r = 2$ and $r = \infty$ respectively; HI – histogram intersection; χ^2 – χ^2 statistics; MD – match distance; KS – Kolmogorov-Smirnov distance; EMD – earth mover's distance.

[1] Maximum values of measures possible to obtain for an eight-bin histogram with unit total mass ($\sum_i h_i = 1$) and maximal ground distance $d = 7/8$.

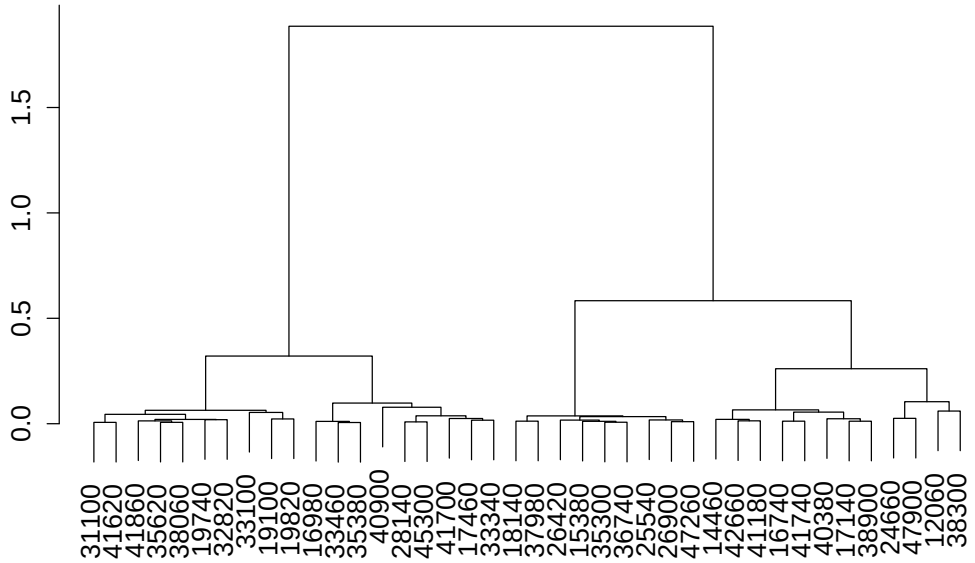


Figure 12: Dendrogram of AHC clustering for 40 MSAs.

7 Clustering of MSAs

This section describes the clustering of MSAs based on their histograms of surrounding ratios. First, we present the clustering process. Next, methods of result visualization are shown. Finally, results of clustering are evaluated.

7.1 Clustering method

The input dataset consists of 40 histograms of MSAs, obtained as described in Section 5.6. Each histogram is composed of 20 bins, discretizing the counterdomain of surrounding ratios into intervals of 5% width. Distances between each pair of histograms are calculated using the Earth Mover's Distance and used to form a distance matrix. Next, *Agglomerative Hierarchical Clustering* (AHC) is performed; see Algorithm 1. Figure 12 shows a dendrogram which is an output of the AHC algorithm. Based on the dendrogram it is possible to divide the dataset in any number of clusters (not greater than the number of objects, here MSAs, of course) and get objects of each cluster.

Algorithm 1 Agglomerative Hierarchical Clustering

1. Initially each object $x_1, x_2, ..., x_n$ is in its own cluster $C_1, C_2, ..., C_n$
2. Repeat until there is only one cluster left:
3. Merge the closest clusters

Table 3 shows assignments, with a number of clusters equal to four. In the first column, MSA identifiers are provided, next the assignments for datasets: 1992 with standard sprawl index, 1992 with intensity sprawl index, 2001 with generalized sprawl index, and so on.

Table 3: Clustering results

MSA	y92std	y92int	y01gen	y01int	y06gen	y06int
12060	1	1	1	1	1	1
14460	1	1	2	2	2	2
15380	2	2	3	2	3	2
16740	1	1	3	1	3	1
16980	3	3	4	3	4	3
17140	1	1	1	1	1	1
17460	3	4	3	4	4	4
18140	2	2	1	4	3	4
19100	4	3	3	2	3	2
19740	4	4	3	2	3	2
19820	4	3	4	3	4	3
24660	1	1	1	1	1	1
25540	2	2	2	4	2	4
26420	2	4	3	2	4	2
26900	2	2	3	4	3	4
28140	3	3	1	1	1	1
31100	4	4	4	3	4	3
32820	4	4	1	4	1	4
33100	4	3	4	3	4	3
33340	3	3	3	2	4	2
33460	3	3	1	2	3	2
35300	2	2	2	2	2	2
35380	3	3	3	2	4	2
35620	4	4	3	2	4	2
36740	2	4	2	1	2	1
37980	2	2	2	4	2	4
38060	4	4	1	2	3	2
38300	1	1	1	1	1	1
38900	1	1	1	2	1	2
40380	1	1	1	1	1	1
40900	3	2	3	2	3	2
41620	4	4	3	3	4	3
41700	3	3	1	1	1	1
41740	1	2	3	3	4	3
41860	4	4	4	3	4	3
42660	1	1	3	2	3	2
41180	1	2	1	4	1	4
45300	3	4	3	2	4	2
47260	2	2	3	4	3	4
47900	1	1	2	1	2	1

Tables 4 and 5 present changes in cluster assignment, first for standard and generalized sprawl index, second for intensity sprawl index. Both tables contain aggregated rows showing the number of occurrences of each sequence of cluster affiliations. For example, the first row of Table 4 shows, that seven MSAs are assigned to the first cluster in all datasets. Another example is the case when an MSA is assigned to the fourth cluster for the 1992 dataset, and to the first one both for 2001 and 2006 datasets, which occurs only once.

Table 4: Changes in cluster assignment (standard and generalized sprawl index)

y92std	y01gen	y06gen	occurrences
1	1	1	7
3	1	1	2
4	1	1	1
1	2	2	2
2	2	2	4
2	1	3	1
3	1	3	1
4	1	3	1
1	3	3	2
2	3	3	3
3	3	3	1
4	3	3	2
1	3	4	1
2	3	4	1
3	3	4	4
4	3	4	2
3	4	4	1
4	4	4	4

Table 5: Changes in cluster assignment (intensity sprawl index)

y92int	y01int	y06int	occurrences
1	1	1	7
3	1	1	2
4	1	1	1
1	2	2	3
2	2	2	3
3	2	2	4
4	2	2	5
2	3	3	1
3	3	3	3
4	3	3	3
2	4	4	6
4	4	4	2

Tables 6 and 7 show changes between datasets from 2001 and 2006 using the generalized sprawl index or intensity sprawl index, respectively. Numbers on the main diagonal indicate no change in assignments, while others show that an MSA is assigned to different clusters for different datasets. Table 6 shows, that there are eleven cases with a change of assignment for

the generalized sprawl index, while Table 7 indicates that there is no change of assignment for the intensity sprawl index.

Table 6: Changes in cluster assignment (generalized sprawl index)

2001 \ 2006	1	2	3	4
1	10	0	3	0
2	0	6	0	0
3	0	0	8	8
4	0	0	0	5

Table 7: Changes in cluster assignment (intensity sprawl index)

2001 \ 2006	1	2	3	4
1	10	0	0	0
2	0	15	0	0
3	0	0	7	0
4	0	0	0	8

Tables 8 and 9 show results of simultaneously clustering datasets from different years. In Table 8 assignments of MSAs for standard and generalized sprawl index are presented. In Table 9 results for intensity sprawl index are shown. Most of the cases from the 1992 dataset are assigned to the first and the second cluster, while cases from the 2001 and 2006 dataset are assigned mostly to the third and the fourth cluster. This shows that the 1992 dataset is different from the datasets from 2001 and 2006 and should not be compared directly.

Table 8: Datasets standard 1992, generalized 2001, and generalized 2006, clustered together

MSA	1992	2001	2006
12060	1	1	1
14460	1	1	1
15380	1	3	3
16740	1	3	3
16980	2	2	2
17140	1	4	4
17460	1	2	2
18140	1	4	4
19100	2	3	3
19740	2	3	3
19820	2	2	2
24660	1	4	4
25540	1	1	1
26420	1	3	3
26900	1	3	3
28140	1	4	4
31100	2	2	2
32820	2	4	4
33100	2	2	2
33340	1	3	3
33460	2	4	4
35300	1	1	1
35380	2	3	3
35620	2	2	2
36740	1	1	1
37980	1	1	1
38060	2	4	4
38300	1	4	4
38900	1	4	4
40380	1	4	4
40900	3	3	3
41620	2	3	3
41700	1	4	4
41740	1	3	3
41860	2	2	2
42660	1	3	3
41180	1	4	4
45300	1	2	2
47260	1	3	3
47900	1	1	1

Table 9: Datasets intensity 1992, 2001, and 2006, clustered together.

MSA	1992	2001	2006
12060	1	4	4
14460	1	3	3
15380	1	3	3
16740	1	4	4
16980	2	1	1
17140	1	4	4
17460	2	4	4
18140	1	4	4
19100	2	3	3
19740	2	3	3
19820	2	1	1
24660	1	4	4
25540	1	4	4
26420	2	3	3
26900	1	4	4
28140	2	4	4
31100	2	2	2
32820	2	4	4
33100	2	1	1
33340	2	3	3
33460	2	3	3
35300	1	3	3
35380	2	3	3
35620	2	3	3
36740	2	4	4
37980	1	4	4
38060	2	3	3
38300	3	4	4
38900	1	3	3
40380	1	4	4
40900	1	3	3
41620	2	1	1
41700	2	4	4
41740	1	1	1
41860	2	1	1
42660	1	3	3
41180	1	4	4
45300	2	3	3
47260	1	4	4
47900	1	4	4

7.2 Evaluation of clustering

To evaluate the quality of clustering, three measures were used: connectivity, silhouette width, and Dunn index. All of them are internal measures, which means that they take as input only the dataset and the clustering partition. The definitions of these measures are presented below.

Connectivity. The first measure, connectivity, shows if neighbors of observations are in the same cluster. Let $nn_{i(j)}$ be the j^{th} nearest neighbor of observation i, and

$$x_{i,nn_{i(j)}} = \begin{cases} 0 & \text{if } nn_{i(j)} \text{ is in the same cluster as } i, \\ 1/j & \text{otherwise.} \end{cases}$$

Then, for a particular clustering partition $\mathcal{C} = \{C_1, \ldots, C_K\}$ of N observations into K disjoint clusters, connectivity is defined as

$$Conn\left(\mathcal{C}\right) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}},$$

where L is a parameter specifying the number of neighbor taken into consideration during the calculation of the connectivity measure. Connectivity takes values between zero and ∞ and should be minimized.

Silhouette width. Desired clusters should be compact and well separated. It means that on the one hand objects are close to each other inside clusters, but on the other they are far away from objects in other clusters. Compactness and separation are contradictory: with an increase of the number of clusters cthe ompactness also increases but separation decreases.

Silhouette width combines compactness and separation in a non-linear way and is defined as

$$S\left(i\right) = \frac{b_i - a_i}{max\left(b_i, a_i\right)},$$

where a_i is the average distance between i and all other observations in the same cluster, and b_i is the average distance between i and the observations in the "nearest neighboring cluster",

$$a_i = \frac{1}{n\left(C\left(i\right)\right)} \sum_{j \in C(i)} dist\left(i,j\right), \quad b_i = min_{C_k \in \mathcal{C} \setminus C(i)} \sum_{j \in C_k} \frac{dist\left(i,j\right)}{n\left(C\left(i\right)\right)},$$

where $C\left(i\right)$ is the cluster containing observation i, $dist\left(i,j\right)$ is the distance between observations i and j, and $n\left(C\right)$ is the cardinality of cluster C. The measure takes values between -1 and 1 and should be maximized.

Dunn index. The Dunn index is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. It is computed as

$$D\left(\mathcal{C}\right) = \frac{min_{C_k, C_l \in \mathcal{C}, C_k \neq C_l}\left(min_{i \in C_k, j \in C_l} dist\left(i,j\right)\right)}{max_{C_m \in \mathcal{C}} diam\left(C_m\right)},$$

where $diam\left(C_m\right)$ is the maximum distance between observations in cluster C_m. The Dunn index takes values between zero and ∞ and should be maximized.

Results

Tables 10 to 15 show evaluation measures for each combination of dataset and sprawl index type. Values of connectivity, Dunn index, and silhouette width are provided for numbers k of clusters varying from 2 to 10.

Table 10: 1992 standard

k	connectivity	Dunn index	silhouette
2	0.33	0.178	0.580
3	1.00	0.187	0.520
4	1.67	0.172	0.474
5	4.00	0.274	0.495
6	6.17	0.341	0.462
7	7.33	0.318	0.419
8	8.83	0.350	0.427
9	10.50	0.350	0.399
10	14.30	0.350	0.405

Table 11: 1992 intensity

k	connectivity	Dunn index	silhouette
2	1.17	0.156	0.542
3	1.17	0.185	0.458
4	5.00	0.218	0.410
5	6.17	0.218	0.418
6	6.17	0.218	0.422
7	8.17	0.218	0.446
8	11.70	0.319	0.426
9	12.20	0.417	0.420
10	15.30	0.389	0.379

Table 12: 2001 generalized

k	connectivity	Dunn index	silhouette
2	2.67	0.088	0.413
3	3.83	0.153	0.389
4	3.83	0.187	0.383
5	5.00	0.188	0.380
6	7.00	0.228	0.376
7	8.50	0.239	0.389
8	10.70	0.266	0.378
9	12.50	0.288	0.370
10	14.70	0.329	0.386

Table 13: 2001 intensity

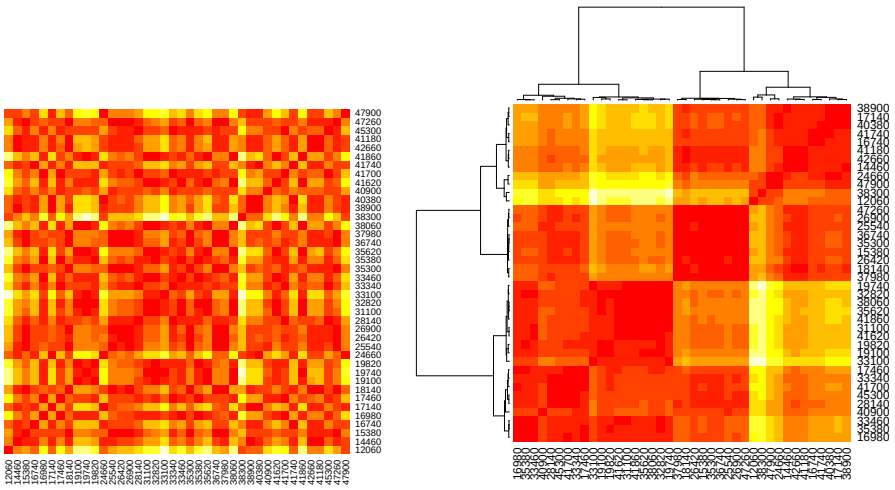
k	connectivity	Dunn index	silhouette
2	1.67	0.109	0.507
3	2.83	0.164	0.475
4	3.17	0.144	0.389
5	5.83	0.259	0.383
6	11.20	0.206	0.331
7	12.00	0.206	0.342
8	13.30	0.234	0.346
9	14.30	0.259	0.328
10	15.00	0.259	0.341

Table 14: 2006 generalized

k	connectivity	Dunn index	silhouette
2	0.00	0.204	0.454
3	3.17	0.129	0.374
4	3.17	0.129	0.398
5	4.33	0.176	0.410
6	7.17	0.177	0.389
7	8.50	0.220	0.381
8	10.50	0.224	0.379
9	12.20	0.224	0.397
10	14.00	0.224	0.381

Table 15: 2006 intensity

k	connectivity	Dunn index	silhouette
2	1.17	0.109	0.513
3	2.83	0.163	0.478
4	4.00	0.126	0.365
5	5.83	0.199	0.349
6	8.00	0.199	0.284
7	9.33	0.253	0.284
8	11.80	0.277	0.270
9	14.70	0.277	0.264
10	16.50	0.309	0.262



(a) unprocessed (b) ordered

Figure 13: Heat maps represent the distance matrix in color scale. Red colors indicate closeness, while light yellow shows large distances between objects. The unordered heat map (a) presents an unprocessed input dataset. The clustered heat map (b) shows results of processing: ordered grid and dendrogram.

7.3 Visualization of clustering results

To visualize both the distance matrix and the results of clustering, three methods are used: cluster heat map, graph display for distance matrices, and multidimensional scaling.

Cluster heat map

A heat map is a grid that presents the similarity between objects using colors. A cluster heat map shows in addition information about the clustering by ordering rows and columns, and adding a dendrogram on two sides of the grid. This way of representing results is very common in biological and biomedical publications, because it is capable to show large amounts of information in a small space [9].

Figure 13 shows heat maps of the previously created distance matrix of MSAs.

Graph display for distance matrices

The second method, graph display for distance matrices, also uses the concept of grid. But in contrast to the heat map it differentiates distances using the size of circles instead of colors. This way of presentation helps to find outliers in a dataset.

Figure 14, presenting the graph display for the analyzed matrix, allows to determine that MSAs 12060, 38300, and 40900 do not fit well to any cluster.

Multidimensional scaling

Multidimensional scaling is a method of presenting a distance matrix in an N-dimensional space. The place of each object is determined such as to keep the distance between all objects as close

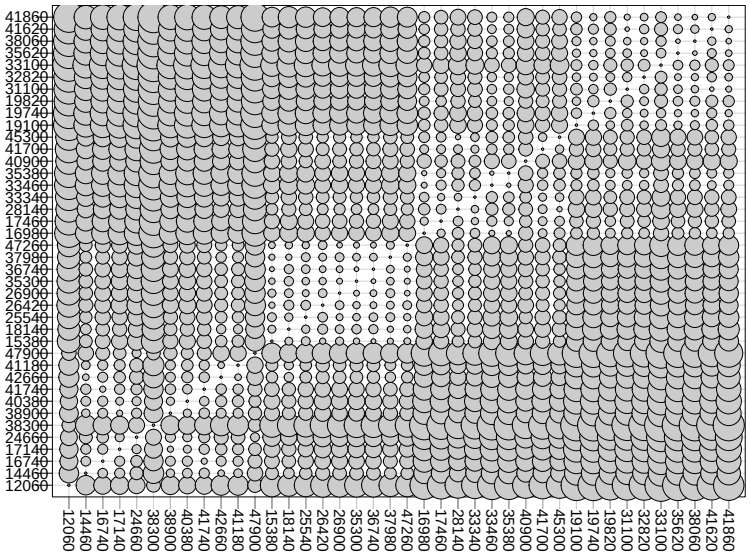


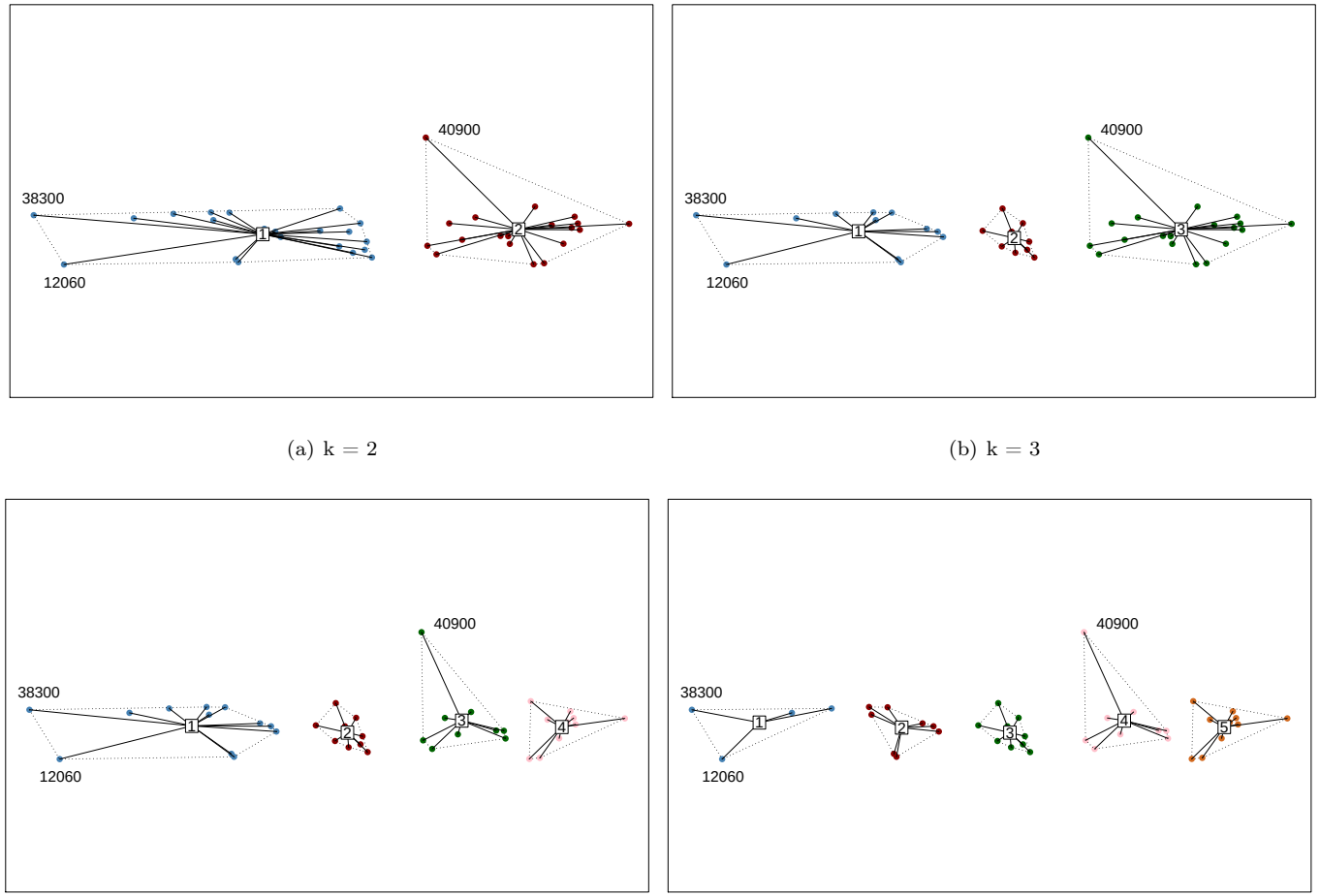
Figure 14: Graph display for distance matrices represents distances by the sizes of circles. This way of visualization facilitates the identification of outliers. An object that does not fit to any cluster is far away from all other objects and therefore all circles in its row or column, except the one on the diagonal, have large sizes (see MSAs 12060, 38300, and 40900).

as possible to the one specified in the input distance matrix. This method of presentation is also suitable to identify outliers.

Figure 15 shows the mapping of the considered distance matrix into a two-dimensional space. MSAs 12060, 38300, and 40900 are far away from the centers of their clusters which confirms the previous observation that they correspond to outliers.

Cluster representatives

Figures 16 to 21 show representatives of the obtained clusters for all datasets and measures. On the left side, average histograms inside clusters are presented. A better way to choose a representative for a cluster is to select the histogram inside the cluster that is closest to the average histogram; this is shown on the right-hand side.



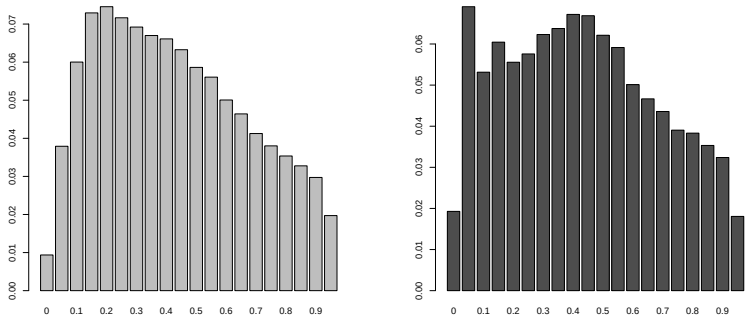
(a) k = 2

(b) k = 3

(c) k = 4

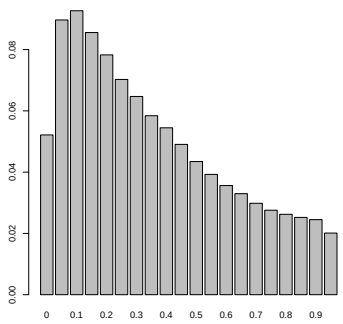
(d) k = 5

Figure 15: Multidimensional scaling presents the distance matrix on a plane. Objects are positioned in a way that distances are as close as possible to those in the distance matrix. Four versions are shown with two (a), three (b), four (c), and five (d) clusters. Three outliers are labeled with their MSA codes.

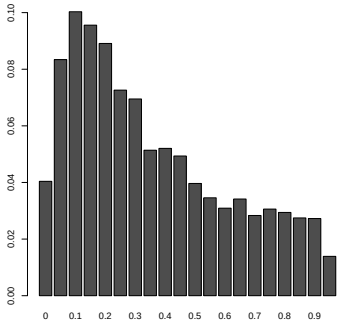


(a) 1992 standard, c1, average

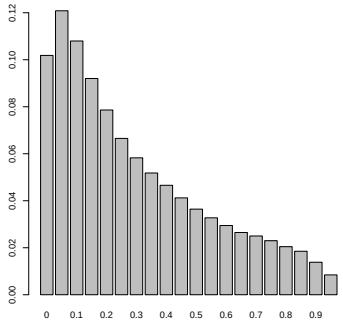
(b) 1992 standard, c1, closest



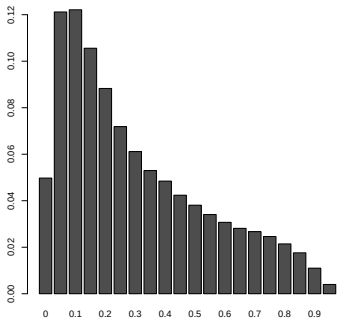
(c) 1992 standard, c2, average



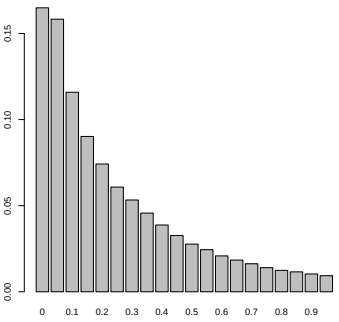
(d) 1992 standard, c2, closest



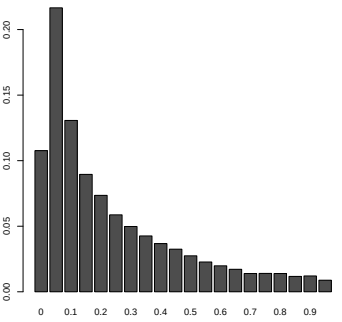
(e) 1992 standard, c3, average



(f) 1992 standard, c3, closest

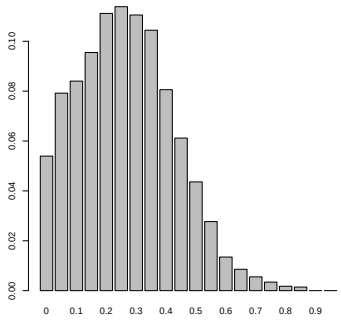


(g) 1992 standard, c4, average

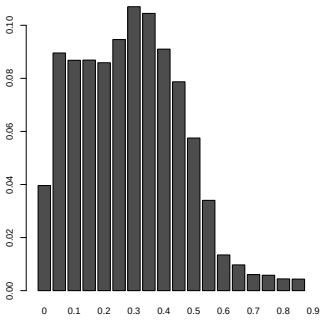


(h) 1992 standard, c4, closest

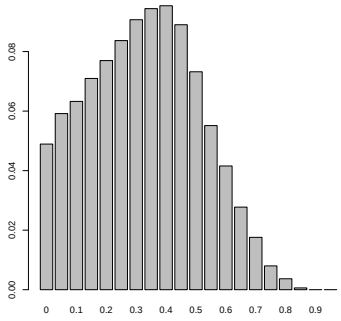
Figure 16: Histograms of cluster representatives, dataset 1992 standard.



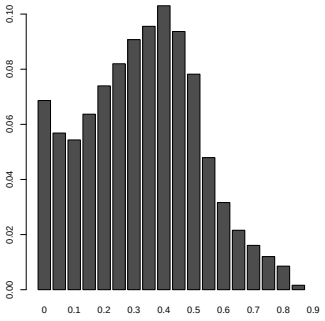
(a) 1992 intensity, c1, average



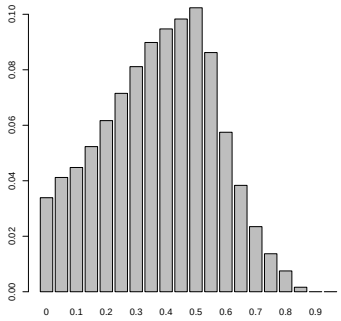
(b) 1992 intensity, c1, closest



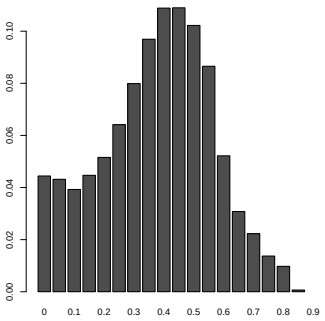
(c) 1992 intensity, c2, average



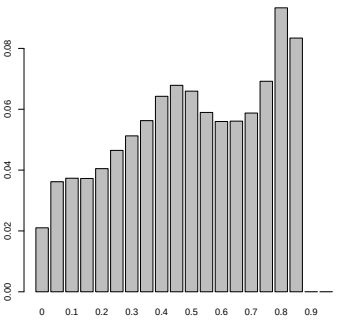
(d) 1992 intensity, c2, closest



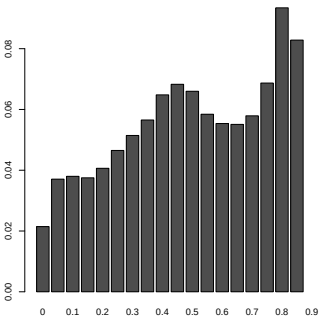
(e) 1992 intensity, c3, average



(f) 1992 intensity, c3, closest

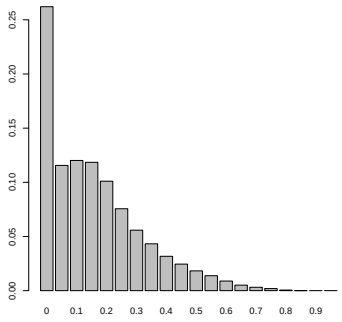


(g) 1992 intensity, c4, average

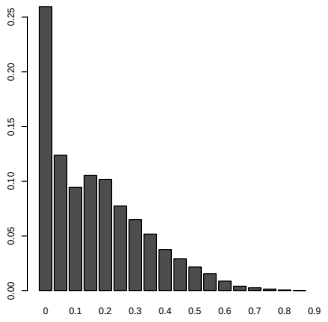


(h) 1992 intensity, c4, closest

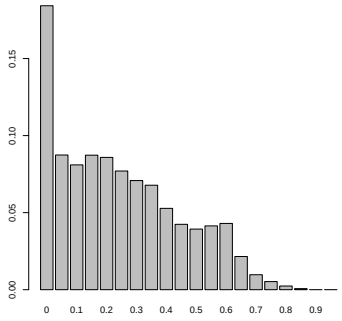
Figure 17: Histograms of cluster representatives, dataset 1992 intensity.



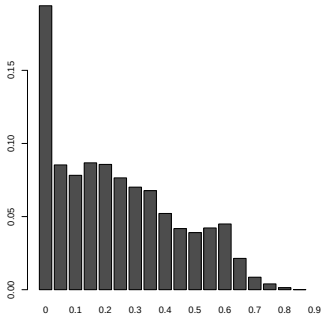
(a) 2001 generalized, c1, average



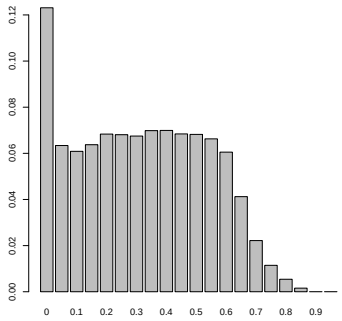
(b) 2001 generalized, c1, closest



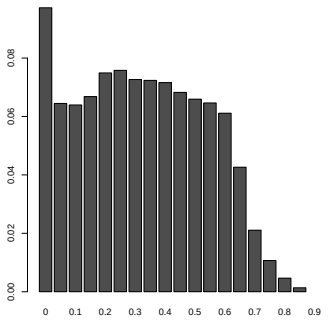
(c) 2001 generalized, c2, average



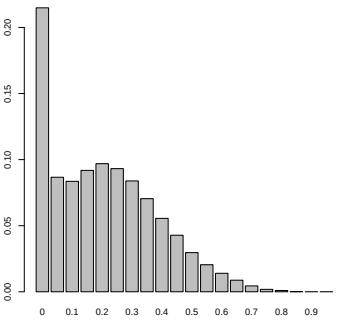
(d) 2001 generalized, c2, closest



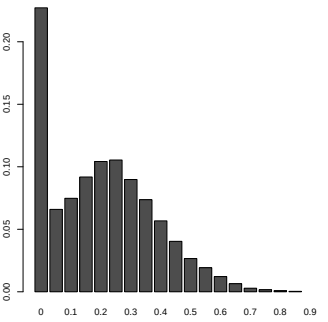
(e) 2001 generalized, c3, average



(f) 2001 generalized, c3, closest

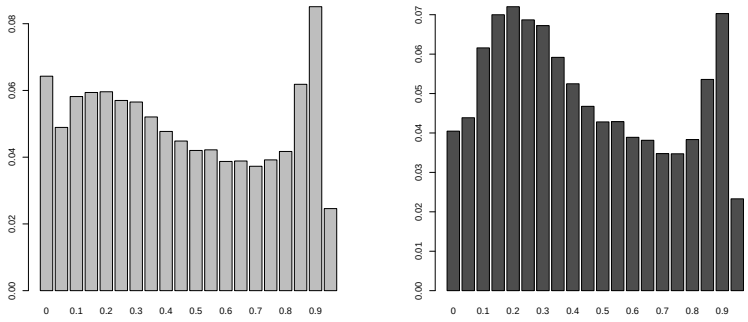


(g) 2001 generalized, c4, average



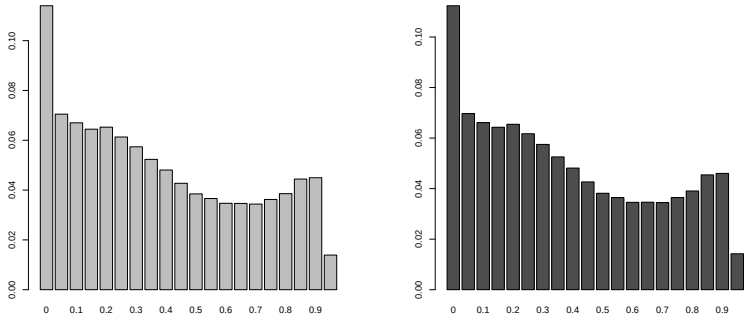
(h) 2001 generalized, c4, closest

Figure 18: Histograms of cluster representatives, dataset 2001 generalized.



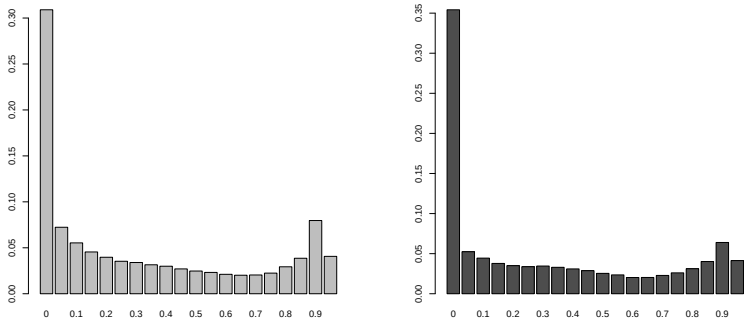
(a) 2001 intensity, c1, average

(b) 2001 intensity, c1, closest



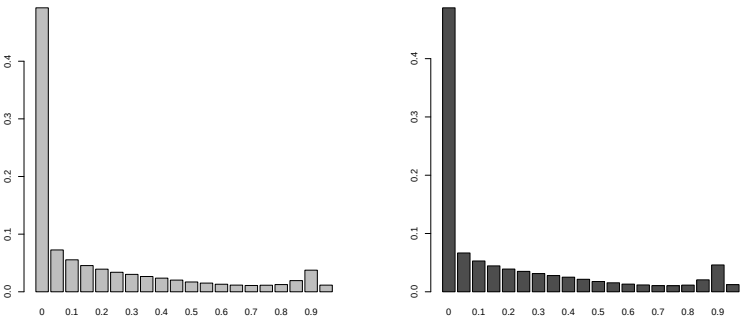
(c) 2001 intensity, c2, average

(d) 2001 intensity, c2, closest



(e) 2001 intensity, c3, average

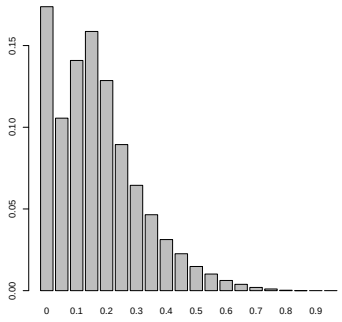
(f) 2001 intensity, c3, closest



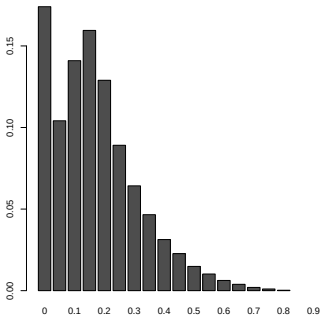
(g) 2001 intensity, c4, average

(h) 2001 intensity, c4, closest

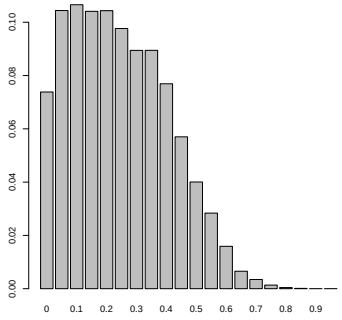
Figure 19: Histograms of cluster representatives, dataset 2001 intensity.



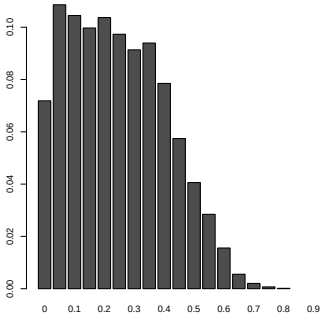
(a) 2006 generalized, c1, average



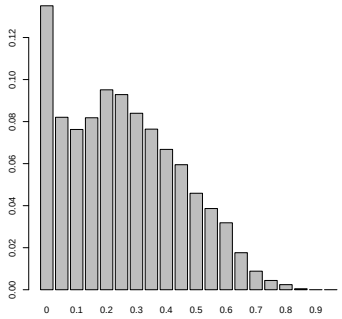
(b) 2006 generalized, c1, closest



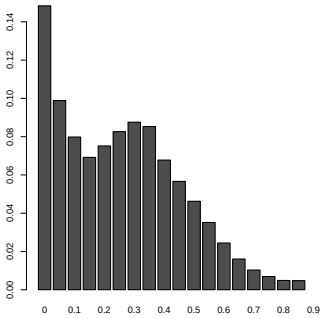
(c) 2006 generalized, c2, average



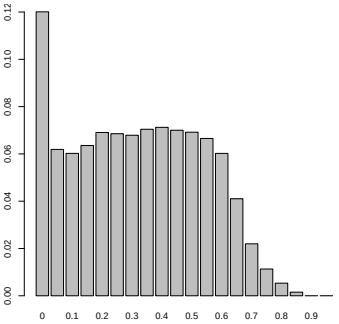
(d) 2006 generalized, c2, closest



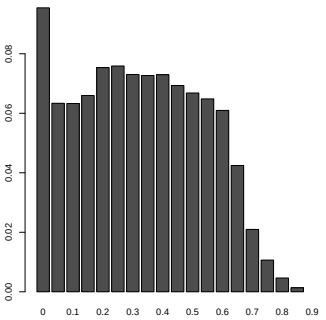
(e) 2006 generalized, c3, average



(f) 2006 generalized, c3, closest

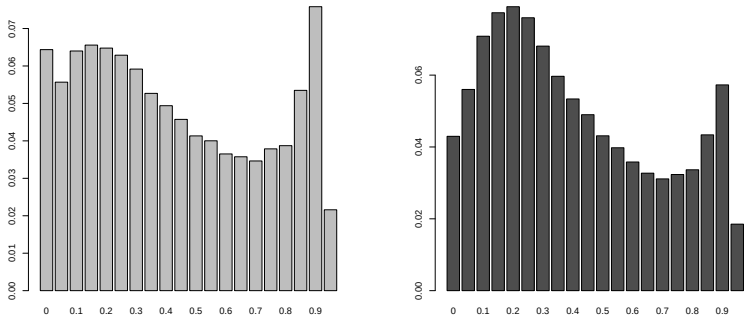


(g) 2006 generalized, c4, average



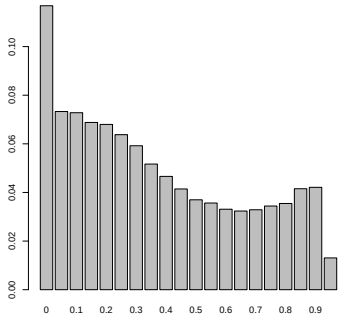
(h) 2006 generalized, c4, closest

Figure 20: Histograms of cluster representatives, dataset 2006 generalized.

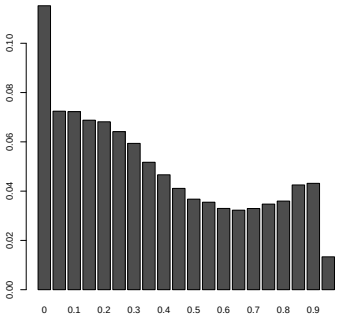


(a) 2006 intensity, c1, average

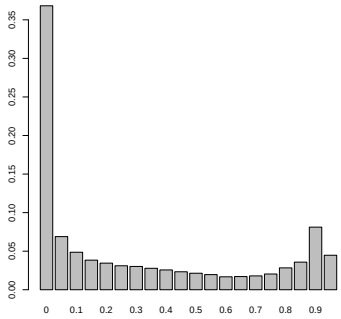
(b) 2006 intensity, c1, closest



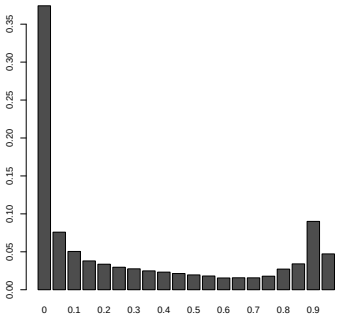
(c) 2006 intensity, c2, average



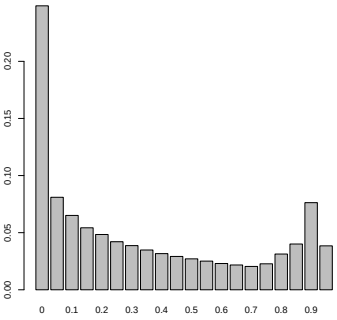
(d) 2006 intensity, c2, closest



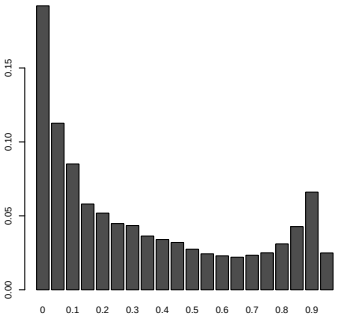
(e) 2006 intensity, c3, average



(f) 2006 intensity, c3, closest



(g) 2006 intensity, c4, average



(h) 2006 intensity, c4, closest

Figure 21: Histograms of cluster representatives, dataset 2006 intensity.

8 Summary

The aim of this report was to describe the process of calculating sprawl indices. First, input datasets were presented. Because different versions of classification system were used in this research, all of them are compared. Next, each step of data preprocessing is described in detail, together with an explanation of the most important fragments of the implemented scripts. In addition, simple examples and mathematical descriptions of the calculations were provided, and the source code used was explained. To adjust the method of calculation to new versions of data and to improve the quality of measuring, some enhancements were proposed. Next, all of them were used to measure sprawl indices of 40 MSAs.

Several histogram dissimilarity measures were tested to choose one suitable for comparing the distribution of surrounding ratios. The Earth Mover's Distance was chosen and applied to calculate distances between distribution of MSAs. Then, the AHC algorithm was used to cluster MSAs with similar distributions of surrounding ratios. Finally, four methods for visualization of clustering results were proposed. They show distances between MSAs, facilitate the identification of outliers, and present characteristics of clusters.

8.1 Future work

The current research gives rise to several directions for future work. These include:

- excluding road networks from the calculation of sprawl indices – suitable shapefiles for this might be: Census 2000 TIGER/Line® Data[1] and National Atlas[2],
- checking the influence of size and shape of the mask used, on the computed sprawl index,
- using other histogram dissimilarity measures and clustering algorithms,
- distinguishing types of undeveloped land
  - water and wetlands as natural barriers,
  - or water and forests as attractive neighborhoods,
- examining the influence of population density
  - is it correlated with sprawl index?
  - for which values sprawl index is highest?
- investigating changes of sprawl over time, taking into account the influence of
  - transportation network,
  - previous landcover,
  - previous population density (locally in neighborhood and globally in a whole MSA).

[1] `http://arcdata.esri.com/data/tiger2000/tiger_download.cfm`
[2] `http://nationalatlas.gov/maplayers.html`

References

[1] Marcy Burchfield, Hanry Overman, Diego Puga, and Matthew Turner. Causes of sprawl: a portrait from space [online]. *London: LSE Research Online*, 2006.

[2] Ross Ihaka and Robert Gentleman. The R project for statistical computing. http://www.r-project.org/, 1993. [Online; accessed 18 June 2013].

[3] An ESRI White Paper. ESRI shapefile technical description. http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf, 1998. [Online; accessed 23 July 2013].

[4] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[5] Gary Sherman. Quantum geographic information system. http://www.qgis.org/, 2002. [Online; accessed 18 June 2013].

[6] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991.

[7] Frank Warmerdam. Geospatial data abstraction library. http://www.gdal.org/, 1993. [Online; accessed 18 June 2013].

[8] Wikipedia. Urban sprawl. URL `http://en.wikipedia.org/wiki/Urban_sprawl`. [Online; accessed 21 October 2013].

[9] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009. doi: 10.1198/tas.2009.0033. URL `http://amstat.tandfonline.com/doi/abs/10.1198/tas.2009.0033`.

A NLCD1992

The classification system used by NLCD1992 is modified from the Anderson Land Cover Classification System.

1 Water
  - 11 Open Water
  - 12 Parennial Ice/Snow

2 Developed
  - 21 Low Intensity Residential
  - 22 Hight Intensity Residential
  - 23 Commercial/Industrial/Transportation

3 Barren
  - 31 Bare Rock/Sand/Clay
  - 32 Quarries/Strip Mines/Gravel Pits
  - 33 Transitional

4 Forest
  - 41 Deciduous Forest
  - 42 Evergreen Forest
  - 43 Mixed Forest

5 Shrubland
  - 51 Shrubland

6 Non-natural woody
  - 61 Orchards/Vineyards/Other

7 Herbaceous Upland
  - 71 Grasslands/Herbaceous

8 Planted/Cultivated
  - 81 Pasture/Hay
  - 82 Row Crops
  - 83 Small Grains
  - 84 Fallow
  - 85 Urban/Recreational Grasses

9 Wetlands
  - 91 Woody Wetlands
  - 92 Emergent Herbaceous Wetlands

B NLCD2001, NLCD2006

The classification system used by NLCD2001 is modified from the Anderson Land Cover Classification System

1 Water
  - 11 Open Water
  - 12 Parennial Ice/Snow

2 Developed
  - 21 Open Space
  - 22 Low Intensity
  - 23 Medium Intensity
  - 24 Hight Intensity

3 Barren
  - 31 Barren Land (Rock/Sand/Clay)

- 4 Forest
  - 41 Deciduous Forest
  - 42 Evergreen Forest
  - 43 Mixed Forest
- 5 Shrubland
  - 51 Dwarf Scrub
  - 52 Shrub/Scrub
- 7 Herbaceous
  - 71 Grasslands/Herbaceous
  - 72 Sedge/Herbaceous
  - 73 Lichens
  - 74 Moss
- 8 Planted/Cultivated
  - 81 Pasture/Hay
  - 82 Cultivated Crops
- 9 Wetlands
  - 91 Woody Wetlands
  - 92 Emergent Herbaceous Wetlands

C Software

This appendix lists main software used for sprawl index computation and the other activities described in this report.

R

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. An important part of R is the set of packages. Below, packages used for computation and visualization are listed.

`http://www.r-project.org/`

package raster

The `raster` package is able to read, write, manipulate, analyze and model gridded spatial data. The package implements basic and high-level functions and processing of very large files is supported.

`http://cran.r-project.org/web/packages/raster/`

package emdist

Package `emdist` provides method for calculation of Earth Mover's Distance.

`http://cran.r-project.org/web/packages/emdist/`

package ade4

Package `ade4` allows to create a graph display representing a distance matrix as a grid of circles using method `table.dist`. The size of the figure depends on the distance between objects.

`http://pbil.univ-lyon1.fr/ADE-4/ade4-html/00Index.html`

package stats

Method `heatmap` from package `stats` enables to create a heatmap presenting a distance matrix.
`http://stat.ethz.ch/R-manual/R-patched/library/stats/html/00Index.html`

package cValid

Package `cValid` provides methods for calculation of the connectivity measure and the Dunn index.
`http://cran.r-project.org/web/packages/clValid/`

GDAL

`GDAL` is a translator library for raster geospatial data formats that is released under an X/MIT style Open Source license by the Open Source Geospatial Foundation. As a library, it presents a single abstract data model to the calling application for all supported formats. It also comes with a variety of useful commandline utilities for data translation and processing. The NEWS page describes the April 2013 GDAL/OGR 1.10.0 release.

preprocessing: ogr2ogr – separation of MSA boundary, gdalwarp – extraction of MSA land cover,
`http://www.gdal.org/`

QGIS

`Quantum GIS (QGIS)` is a user friendly Open Source Geographic Information System (GIS) licensed under the GNU General Public License. QGIS is an official project of the Open Source Geospatial Foundation (OSGeo). It runs on Linux, Unix, Mac OSX, Windows and Android and supports numerous vector, raster, and database formats and functionalities.

preprocessing: QgsGeometryAnalyzer – buffering MSA boundaries.
`http://www.qgis.org/`

D Scripts

In this appendix implemented scripts are shortly described.

model.r

Script `model.r` performs linear regression of Sprawl index based on socio-economical factors (as proposed by Burchfield et al. in [1]).

clustering.r

Script `clustering.r` performs *Agglomerative Hierarchical Clustering* (AHC) on chosen dataset (result file with distribution of surrounding ratio). *Earth Mover's Distance* implemented in package `emdist` is used to calculate measure of dissimilarity between distributions. Script visualizes calculated distance matrix and results of clustering (distance table, heat map, classical multidimensional scaling). It evaluates quality of clustering using internal measures (connectivity, Dunn index, silhouette). Cluster representatives (average histogram and closest real instance) are determined.

clustering_compare.r

Script `clustering_compare.r` compares results of clustering using different datasets. Comparison is shown in form of table. Also, latex code is produced to facilitate export.

sprawl_map.r

For chosen dataset script `sprawl_map.r` plots map of MSA with information about sprawl index included in color scale. Data are loaded from result file with distribution of surrounding ratio of each MSA.

population_density.r

Script `population_density.r` reprojects dataset with population density and plots data together with boundaries of MSA. It offers also possibility to filter and plot only areas with specified population density.

histogram_distances.r

Script `histogram_distances.r` compares histograms. It plots them and computes values of different histogram distances.

clustering_socio_economical.r

Script `clustering_socio_economical.r` performs clustering of MSAs based on socio-economical factors. Clustering based on data used in [1] for building linear model.

buffer.py

Script `buffer.py` buffers specified shapefile and saves result to new file. Size of buffer is specified by user.

removeOldTmp.sh

Script `removeOldTmp.sh` removes old temporary files. It is used during computation of Sprawl index to neutralize the risk of filling Swap space.

plot_histogram.r

Script `plot_histogram.r` plots histograms (standard and cumulative) for all/selected MSA.

selectedMSA.r, selectedMSA_92.r

Scripts `selectedMSA.r` and `selectedMSA_92.r` for each of specified MSA perform preprocessing steps and execute `sprawlIndexBatch.r` script.

sprawlIndexBatch.r

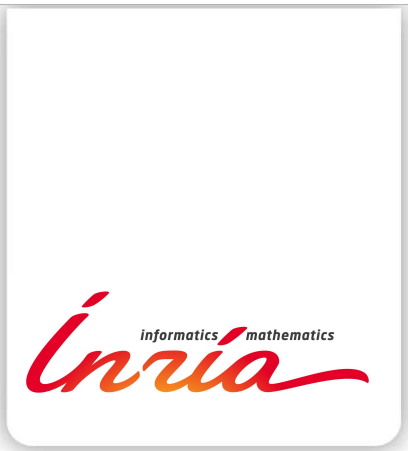
Script `sprawlIndexBatch.r` loads raster file, defines mask and executes `sprawlIndex.r` script.

sprawlIndex.r

Script `sprawlIndex.r` performs computation of Sprawl index, plots maps and histogram, and save results to csv file.

clustering_library.r

Script `clustering_library.r` creates distance matrix for specified list of distributions using Earth Mover's Distance.



RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399